様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

令和　５ 年　 ５ 月 １０ 日現在

機関番号：１２６０１
研究種目：挑戦的研究（萌芽）
研究期間：2020～2022
課題番号：２０Ｋ２１８２７

研究課題名（和文）グラフ匿名化による大規模個人ゲノムデータベースの安全で効率的な検索技術の開拓

研究課題名（英文）Effcient and secure search via graph annymization techniques for large individual genome database

研究代表者
渋谷　哲朗（Shibuya, Tetsuo）

東京大学・医科学研究所・教授

研究者番号：６０３９６８９３

交付決定額（研究期間全体）：（直接経費）　　4,800,000円

研究成果の概要（和文）：χ2統計量やコクラン・アーミテージ統計量、TDT統計量、統計上位遺伝子など、ゲノムワイド関連解析における基盤統計値の公開を差分プライバシーの観点からプライバシー保護する一連の手法の開発に成功した。さらに、差分プライバシーとk-匿名化の技術の統合、関連グラフ理論の構築、グラフ理論に基づくゲノム圧縮技術の開発、多様なデータからのグラフ情報抽出のための基盤技術の開発、データ保護技術の開発いも成功した。

研究成果の学術的意義や社会的意義
これらの研究によって公開が可能となったゲノムワイド関連解析における多くの統計値は、大規模個人ゲノム解析におけるもっとも基盤的な重要データであり、これらをプライバシーの保護を図りつつ公開できるようになったことは非常にインパクトのある成果である。特に統計上位遺伝子の公開に関する研究は、プライバシー保護分野のトップ国際会議においてIEEE Outstanding Paper Awardを受賞するなど、国際的にも高い評価を得た。

研究成果の概要（英文）：We developed differential privacy-based new technologies to publish various statistics of genome wide association study (GWAS), such as Chi-square statistic, Cochran-Armitage trend test, TDT test, and the k-top associated genes. We also developed a method of combining differential privacy and k-anonymization, related graph theory, a graph-based genome compression algorithm, graph feature extraction methods for various databases, data protection methods.

研究分野：バイオインフォマティクス・アルゴリズム

キーワード：アルゴリズム理論　プライバシー保護　差分プライバシー　データ保護　個人ゲノム

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９（共通）

１．研究開始当初の背景

次世代シークエンサーとよばれるゲノム解読技術の革命的進展により、ゲノム情報をきわめて安価・高速に得ることができるようになった。これらの情報の管理には、プライバシー保護の観点から極めて高い安全性が要求される一方で、そのデータ量も非常に大きく、近い将来に日本全国民のデータが管理された状態を想定すると、そのデータ量は 10 エクサバイトにも達すると考えられている。このため、そのような超大規模データへのスケーラビリティと、高い安全性を両立するプライバシー保護技術の開発が求められている。

これまで、様々なプライバシー保護技術の研究が行われ、新しい画期的な技術がいくつか登場している。ひとつは完全準同型演算などに代表される秘匿計算とよばれる手法であるが、これらはきわめて低速であり、全国民規模も考える必要のある大規模ゲノムデータベースに適用するには現状の技術ではきわめて困難である。別の手法として、k-匿名化などに代表される差分プライバシー技術がある。これは、データベースのデータにノイズなどを加えるなどの加工を行うことで個人のプライバシーを保証する。これらの手法の中には秘匿計算などと比較すれば効率性、スケーラビリティの高い技術も存在する。しかし、一般的にゲノム研究は研究進展を望む患者の個人データを用いて行う研究であり、提供されたデータを最大限活用して最大限の成果を上げることが倫理的に求められる点で、通常の個人情報データと性質を異とする。そのような中、ノイズ加工等によってデータを棄損することはデータ提供者である患者の要望に沿ったものと倫理的に言えるかどうかは大きな疑問である。

このような背景から、今後の個人ゲノム時代の超大規模ゲノムデータベースを活用して医科学をさらに発展させるにあたっては、個人ゲノムデータの倫理的特性にあった高いレベルの個人情報保護と、効率性・スケーラビリティを両立する検索・解析技術を実現することがきわめて喫緊の課題である。しかし、現状ではこれを解決する一般的なプライバシー保護技術は存在しないため、ゲノム情報という問題の性質を最大限活用して解決する必要がある。これに対しゲノムグラフと呼ばれるゲノムに特化したデータ構造を活用すれば、この問題を解決される可能性がある。

２．研究の目的

次世代シークエンサーとよばれるゲノム解読技術の革命的進展により、日本を含め世界各国で数十万人規模の大規模なゲノムの収集プロジェクトが行われ、患者ひとりひとりに対するゲノム医療や、D2C(Direct to Consumer)サービスなどの試みも活発になされるようになった。　このゲノム情報は究極の個人情報とも称せられ、それらを研究や診療に用いるには提供者の承諾や厳格な倫理審査の上、非常に高いレベルのセキュリティで管理する必要がある。これに対し、近年、いくつかの画期的なプライバシー保護技術が登場している。ひとつは完全準同型演算などに代表される秘匿計算とよばれる手法で、これらはデータを暗号化したまま演算を行うことができる。しかし秘匿計算手法の多くはきわめて低速であり、究極的には全国民規模も考えられるゲノムデータベースへの適用は現状では極めて困難である。一方、より効率性、スケーラビリティの高いプライバシー保護技術として、データベースのデータにノイズを加えるなどの加工を行うことで個人のプライバシーを保証する差分プライバシー技術も活発に研究されている。しかし、一般的にゲノム研究は善意の提供者個人の究極のデータを用いて行う研究であり、倫理上患者の意思・要望を尊重する必要がある一方で、データ提供者である多くの患者は提供データを最大限活用して最大限の成果を上げることを希望しているともされる。たとえ性能差が小さくとも、ノイズ等の加工を加えて解析を行うことがデータ提供者である患者の要望に沿ったものと倫理的に言えるかどうかは大きな問題である。

本研究の目的は、今後の個人ゲノム時代の超大規模ゲノムデータベースを活用して医科学をさらに発展させるため、大規模個人ゲノムデータベースに対し、究極の個人情報ともいわれるゲノムデータの倫理的特性にあった安全かつ効率的な検索・解析手法を実現することである。この実現には、データを棄損しない効率的な匿名化技術の開発が必要である。本研究計画では、様々なデータを複雑に表現できるグラフ構造を用いてゲノムを扱うことによって、必要なデータを保持しつつ匿名化が達成できる手法の確立をめざす（グラフ匿名化）。特に、複数のゲノムをコンパクトに表現する技術として、ゲノムグラフとよばれる技術があり、本研究では、このゲノムグラフを発展・拡張することによってこれを実現することをめざした。さらに、当該技術を将来の個人ゲノム研究の基盤として開拓することを狙い、そのグラフ匿名化技術によって匿名化されたデータに対する検索・解析技術基盤の開拓も狙った。

３．研究の方法

本研究は、大規模ゲノム情報に対するグラフ匿名化技術の開拓とグラフ匿名化データに対する

検索・解析技術の開発の開拓の二つの柱からなった。これに加え、それによるゲノム匿名化基盤の実現に必要な周辺技術の開発を同時に進めていった。

まず、大規模ゲノム情報を匿名化するにあたって、複雑な情報を表現可能なグラフを活用する技術を確立を目指した。グラフは様々な要素の関係を複雑に記述できる一方で、含まれる部分グラフを探索することが計算理論的に困難（NP-困難）であるなど、情報論的な秘匿技法とできる可能性があった。また、ゲノムグラフとよばれるグラフは次世代シークエンサー出力等の情報からゲノムを再構築するゲノム・アセンブリとよばれる情報処理にも用いられる技法であるが、このデータ構造は配列全体を再構築できるにもかかわらず、配列の由来は保持していない、という性質を持つ。このゲノムグラフのようなグラフ構造をうまく発展・応用することができればゲノム配列の匿名化に用いることができる可能性がある。そこで本研究では、このゲノムグラフを拡張・発展させグラフ匿名化を実現する可能性を検討した。同時にその他様々なグラフ表現も検討し、グラフ匿名化の実現を狙うとともに、秘匿性能の検証を行っていった。

次に、ゲノムデータベースは大規模であるため、グラフ匿名化されたゲノムデータに対しても効率的な検索や解析が可能である必要がある。そのため、ゲノム匿名化されたデータベースに対する検索技術や解析技術の開発を検討した。特に、これまでゲノムグラフを用いた様々な配列検索や疾患遺伝子解析など配列解析のためのアルゴリズムが研究されており、それらのアルゴリズムを拡張・発展させ匿名化データに対応することを軸に開発を進めた。

さらに、プライバシー保護技術は、匿名化、暗号化、アクセス秘匿、クエリー秘匿、データ利用者秘匿、など多岐にわたり、ゲノムデータの保護は、グラフ匿名化だけによるのではなく、他の様々な技術と合わせてより安全に研究を遂行できる環境を構築する必要がある。そのため、グラフ匿名化とあわせて用いることのできる他のプライバシー保護技術についても、検討を行い、ゲノムの匿名化技術の基盤構築をめざした。

４．研究成果

（１） 先進的匿名化技術の開発

データベース公開の際のプライバシー保護技術として、χ2 統計量やコクラン・アーミテージ統計量などゲノムワイド関連解析における多くの基盤的統計値に対し差分プライバシーを活用したゲノム統計値公開手法を開発することに成功した(Bioinfo. Adv. 2021)。また TDT 統計量の公開を高速に計算する手法の開発にも成功した(BIBM 2021, PSB 2022)。また、それらの統計値から上位のものを公開する際のより高精度な公開手法の開発にも成功した(IEEE TrustCom 2022)。

これらの研究によって公開が可能となったゲノムワイド関連解析における多くの統計値は、大規模個人ゲノム解析におけるもっとも基盤的な重要データであり、これらをプライバシーの保護を図りつつ公開できるようになったことは非常にインパクトのある成果である（図２）。特に統計上位遺伝子の公開に関する研究（IEEE TrustCom 2022）は、プライバシー保護分野のトッ

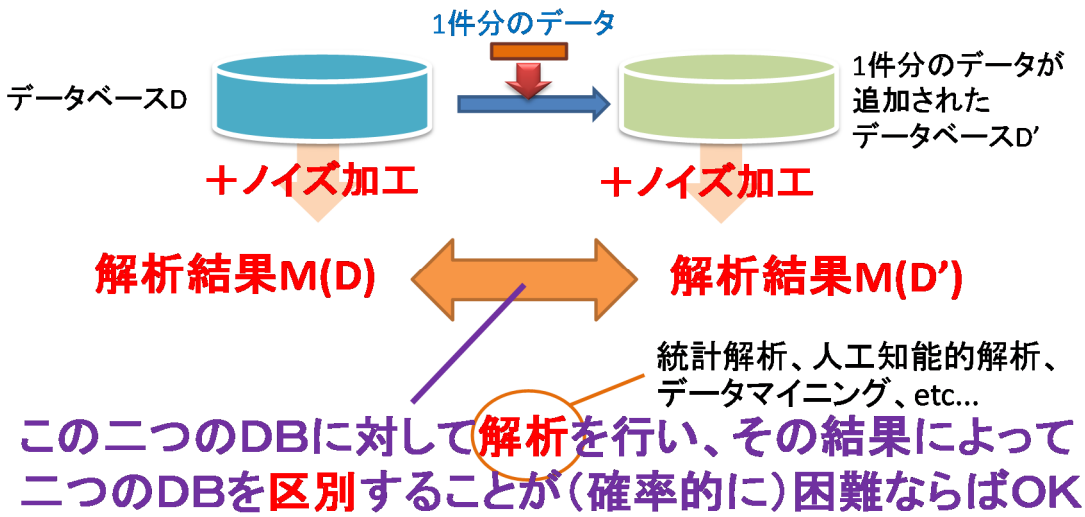

図１　差分プライバシー

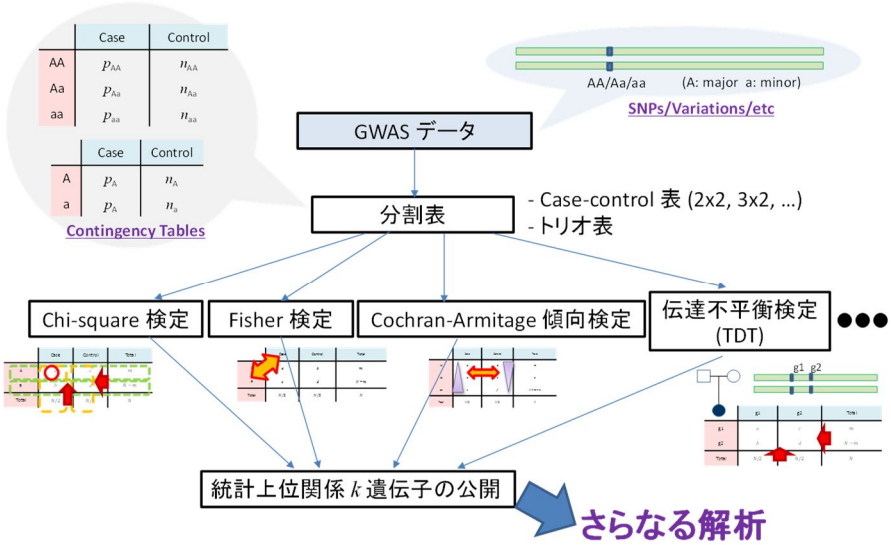

図２　ゲノムワイド関連解析における重要基盤統計量

プ国際会議において IEEE Outstanding Paper Award を受賞するなど、国際的にも高い評価を得た。

さらに k-匿名化と差分プライバシーの双方の要請に対応する新たな匿名化技術の開発にも成功した(HEALTHINF 2023)。これは差分プライバシーといわゆる k-匿名化（図３参照）が守りたいプライバシーの性質が異なることに着目し、その双方を保護する全く新しい手法である。

（２） グラフ理論に基づくゲノム圧縮技術の開発

ゲノムグラフの新たな効率的表現をグラフ理論を活用して圧縮表現する技術を開発、実装を行った（WABI 2021)。

（３） 関連グラフ理論の構築

グラフ匿名化の基盤となるグラフ理論の研究として、グラフ匿名化の際のグラフ情報の編集がどのようにグラフの特質に影響を与えるか、について理論的な結果を得ることに成功した(COCOON 2022, CCCG 2022, SOFSEM 2023)。

氏名	生年月日	郵便番号	性別	さまざまな情報
白金華子	19990123	108-8639	男	...
駒場一朗	19990711	153-8902	男	...

危ない情報

仮名	生年月日	郵便番号	性別	さまざまな情報
PB924CD	1999****	1**-8***	男	...
AR325HB	1999****	1**-8***	男	...

ここに関してはk=2になった

図３　k-匿名化

さらに文字列は一種のグラフとしてとらえることもできるが、文字列データに対する編集についても文字列の特質にどのような影響を与えるかについての理論的な結果を得ることにも成功した(ICS 2022)。

（４）多様なデータベースからのグラフ情報の抽出技術の開発

医療情報の多くが自然言語で記述されることから、それらの情報のプライバシー保護を行うための基盤となる自然言語をグラフ的に理解するための重要な支援言語技術の開発に成功した（CLEF 2020, ICMLA 2020, CCBD 2021, LREC 2022)。
また、がんの遺伝子発現情報からグラフ情報を抽出するための基盤技術の開発にも成功した(BMC Bioinfo. 2023)。

（５）データ保護技術の開発

グラフ理論的な観点を活用し、データ保護を行う wear leveling 技術の開発にも成功した(ISAAC 2020)。

５．主な発表論文等

〔雑誌論文〕　計12件（うち査読付論文　12件／うち国際共著　2件／うちオープンアクセス　6件）

１．著者名 Yamamoto Akito、Shibuya Tetsuo	４．巻 1-1-vbab004
２．論文標題 More practical differentially private publication of key statistics in GWAS	５．発行年 2021年
３．雑誌名 Bioinformatics Advances	６．最初と最後の頁 1-10
掲載論文のDOI（デジタルオブジェクト識別子） 10.1093/bioadv/vbab004	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

１．著者名 Kazushi Kitaya and Tetsuo Shibuya	４．巻 201(12)
２．論文標題 Compression of Multiple k-mer Sets by Iterative SPSS Decomposition	５．発行年 2021年
３．雑誌名 Proc. WABI 2021, Leibniz International Proceedings in Informatics (LIPIcs)	６．最初と最後の頁 1-12
掲載論文のDOI（デジタルオブジェクト識別子） 10.4230/LIPIcs.WABI.2021.12	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

１．著者名 Robert Barish and Tetsuo Shibuya	４．巻 1
２．論文標題 Solving teleportation mazes with limited visibility	５．発行年 2021年
３．雑誌名 The 23rd Thailand-Japan Conference on Discrete and Computational Geometry, Graphs, and Games	６．最初と最後の頁 110-111
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

１．著者名 Yamamoto Akito、Shibuya Tetsuo	４．巻 1
２．論文標題 Differentially Private Linkage Analysis with TDT ? the case of two affected children per family	５．発行年 2021年
３．雑誌名 IEEE International Conference on Bioinformatics and Biomedicine	６．最初と最後の頁 765-770
掲載論文のDOI（デジタルオブジェクト識別子） 10.1109/bibm52615.2021.9669365	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

１．著者名 Akdemir Arda、Shibuya Tetsuo	４．巻 1
２．論文標題 UDON: Unsupervised Data SelectiON for Biomedical Entity Recognition	５．発行年 2021年
３．雑誌名 Proc. 4th International Conference on Computing and Big Data	６．最初と最後の頁 1-7
掲載論文のDOI（デジタルオブジェクト識別子） 10.1145/3507524.3507525	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

１．著者名 Akito Yamamoto, Tetsuo Shibuya	４．巻 27
２．論文標題 Efficient Differentially Private Methods for a Transmission Disequilibrium Test in Genome Wide Association Studies	５．発行年 2022年
３．雑誌名 Pacific Symposium on Biocomputing	６．最初と最後の頁 85-96
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

１．著者名 Akdemir Arda、Shibuya Tetsuo、Gungur Tunga	４．巻 1395
２．論文標題 A Comprehensive Analysis of Subword Contextual Embeddings for Languages with Rich Morphology	５．発行年 2021年
３．雑誌名 Advances in Intelligent Systems and Computing	６．最初と最後の頁 31～72
掲載論文のDOI（デジタルオブジェクト識別子） 10.1007/978-981-16-3357-7_2	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

１．著者名 Akito Yamamoto, Tetsuo Shibuya	４．巻 30(2)
２．論文標題 Privacy-Preserving Statistical Analysis of Genomic Data using Compressive Mechanism with Haar Wavelet Transform	５．発行年 2023年
３．雑誌名 Journal of Computational Biology	６．最初と最後の頁 176-188
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

1．著者名 Arda Akdemir, Yeojoo Jeon and Tetsuo Shibuya	4．巻 13
2．論文標題 Developing Language Resources and NLP Tools for the North Korean Language	5．発行年 2022年
3．雑誌名 Conference on Language Resources and Evaluation	6．最初と最後の頁 5595-5600
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

1．著者名 Arda Akdemir, Tetsuo Shibuya	4．巻 2696
2．論文標題 Transfer Learning for Biomedical Question Answering	5．発行年 2020年
3．雑誌名 Proc. CLEF 2020, CEUR Workshop Proceedings	6．最初と最後の頁 1-15
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

1．著者名 Taku Onodera, Tetsuo Shibuya	4．巻 181(65)
2．論文標題 Wear Leveling Revisited	5．発行年 2020年
3．雑誌名 Leibniz International Proceedings in Informatics (LIPIcs)	6．最初と最後の頁 1-17
掲載論文のDOI（デジタルオブジェクト識別子） 10.4230/LIPIcs.ISAAC.2020.65	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 該当する

1．著者名 Arda Akdemir, Tetsuo Shibuya, Tunga Gungor	4．巻 1
2．論文標題 Subword Contextual Embeddings for Languages with Rich Morphology	5．発行年 2020年
3．雑誌名 Proc. ICMLA 2020	6．最初と最後の頁 994-1001
掲載論文のDOI（デジタルオブジェクト識別子） 10.1109/ICMLA51294.2020.00161	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 該当する

〔学会発表〕　計4件（うち招待講演　2件／うち国際学会　2件）

１．発表者名
Akito Yamamoto, Tetsuo Shibuya

２．発表標題
Efficient Differentially Private Methods for a Transmission Disequilibrium Test

３．学会等名
第６７回バイオ情報学研究会

４．発表年
2021年

１．発表者名
Arda Akdemir, Tetsuo Shibuya

２．発表標題
UDON: Unsupervised Data SelectiON for Biomedical Entity Recognition

３．学会等名
The 27th East Asia Joint Symposium（国際学会）

４．発表年
2021年

１．発表者名
渋谷哲朗

２．発表標題
スパコンシステム「SHIROKANE」とゲノム医療

３．学会等名
第9回生命医薬情報連合大会（招待講演）

４．発表年
2020年

１．発表者名
Tetsuo Shibuya

２．発表標題
Algorithmic Challenges for Biomedical Big Data

３．学会等名
The 11th International Conference on Bioscience, Biochemistry and Bioinformatics（招待講演）（国際学会）

４．発表年
2021年

〔図書〕　計0件

〔産業財産権〕

〔その他〕

―

６．研究組織

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考

７．科研費を使用して開催した国際研究集会

〔国際研究集会〕　計0件

８．本研究に関連して実施した国際共同研究の実施状況

共同研究相手国	相手方研究機関